

2.0 Basic distributional concepts

Consider a measurement made on each member of a population of finite size.
outcome set.

The *population distribution* can then be described by listing each value in the outcome set along with the frequency with which members of the population take that value.

For example, consider the hourly wages of full-time white women workers in the U.S. in 1998, measured to the closest penny. The distribution is the list of each value the wage takes (e.g., \$0.00, \$0.01, \$0.02, ...) along with the number of women with that wage.

The *relative frequency distribution* replaces the frequency with the relative frequency (i.e., proportion) of members taking the value.

Probability Mass Function

Let X denote the value for a member of the population selected at random from the population. Then X is a random variable taking on values from the outcome set with probability given by the corresponding relative frequency.

In this case X is a *discrete* random variable as it takes on only a finite number of possible values.

The *probability mass function* of X is then a listing of each value x , say, in the outcome set along with the probability that X takes on the value.

We will denote this number by $P(X = x)$ for each x (in words, “the probability that $X = x$ ”).

Note that we will always have:

$$0 \leq P(X = x) \leq 1 \quad \text{for any } x$$

with the function strictly positive for values in the outcome set, and

$$\sum_x P(X = x) = 1$$

where the sum is over the outcome set.

Ex. The figure is a graph of the probability mass function for women’s annual earnings, based on 1998 March Current Population Survey (CPS).

In many situations it may be desirable to approximate the probability mass function of X by using a mathematically tractable or conceptually simpler form.

For example, in the above graph we have placed a smooth curve through $P(X = x)$ and could use it to describe the distribution of earnings.

Such approximations allow us to summarize the main features of the distribution using a continuous function even when the underlying probability mass function is discrete.

Other examples are *histograms* and the *normal probability curve*. The latter is a parametric approximation that leads to great parsimony if it is accurate.

Probability Density Function

The continuous analog of the probability mass function – a *probability density function* (PDF) – to describe the distribution of probability over the outcome set. The PDF is a function $f(x)$ where x is in the outcome set, such that:

$$f(x) \geq 0 \quad \text{for all } x$$

$$\int_{-\infty}^{\infty} f(x) dx = 1.$$

The PDF enables probabilities to be calculated using the relationship:

$$P(a \leq X \leq b) = \int_a^b f(x) dx \quad a \leq b.$$

Two continuous distributions are worth noting here

The *uniform distribution* on the outcome space the interval $[0,1]$, and is defined by the PDF:

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

The *standard normal distribution*, which has outcome space the set of all real numbers on the interval $(-\infty, \infty)$, and is defined by the PDF:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad -\infty < x < \infty.$$

The graph of this PDF is often called the “bell curve,” and is the most common distributional approximation used in statistical methods.

Cumulative Distribution Function

A distribution, whether continuous or discrete, can also be characterized by its cumulative distribution function (CDF):

$$F(x) = P(X \leq x) \quad \text{for each } x \text{ in the outcome space.}$$

That is $F(x)$ gives the probability that a randomly chosen value is less than or equal to x . If X is discrete we have

$$F(x) = \sum_{y \leq x} P(X = y) \quad \text{for each } x \text{ in the outcome space,}$$

and if X is continuous

$$F(x) = \int_{-\infty}^x f(y) dy \quad \text{for each } x \text{ in the outcome space}$$

These relationships can be inverted to express the PDF in terms of the CDF. In the discrete case, this is

$$P(X = x) = F(x) - F(x-),$$

where x is in the outcome space and $x-$ is the largest value in the outcome space smaller than x . In the continuous case, the relationship is:

$$f(x) \equiv \frac{d}{dx} F(x) \equiv \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h}. \quad (2:1)$$

Quantile Function

A useful quantity related to the CDF is the *inverse cumulative distribution function*, also called the *quantile function*.

$$Q(p) = F^{-1}(p) = \inf_x \{x \mid F(x) \geq p\}.$$

The quantile $Q(p)$ can be thought of as the value of y below which a proportion p of the values fall.

One can also say that this value defines the p th quantile of the population (or equivalently, of the probability distribution of X).

Special cases are the *median* ($p = 0.5$) and the lower and upper *quartiles* ($p = 0.25, p = 0.75$, respectively).

Two common ways to express the quantile function are through *deciles* (i.e., the quantiles corresponding to 0.0, 0.1, ..., 0.9, 1.0) and *percentiles* (i.e., the quantiles corresponding to 0.00, 0.01, ..., 0.99, 1.00).

For example, the bottom decile is the quantile corresponding to $p = 0.10$.

In the earnings distribution, the bottom decile is $Q(0.1) = \$11,500$. The median and upper quartiles are $Q(0.5) = \$24,000$ and $Q(0.75) = \$34,000$, respectively.